Multi-platform Next-generation Sequencing Identifies Novel RNA Molecules and Transcript Isoforms of the Endogenous Retrovirus Isolated from Cultured Cells

One-sentence summary: The authors used short- and long-read RNA sequencing techniques along with PCR analysis to uncover a complex transcriptional landscape in the porcine endogenous retrovirus (PERV).

Norbert Moldován[1], Attila Szűcs[1], Dóra Tombácz[1,2], Zsolt Balázs[1], Zsolt Csabai[1], Michael Snyder[2], Zsolt Boldogkői[1*]

[1]Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary

[2]Department of Genetics, School of Medicine, Stanford University, Stanford, California, USA

Keywords: RNA-Seq, full-length sequencing, transcriptome, Pacific Biosciences, Oxford Nanopore Technologies, endogenous retrovirus

Abstract

In this study, we applied short- and long-read RNA sequencing techniques, as well as PCR analysis to investigate the transcriptome of the porcine endogenous retrovirus (PERV) expressed from cultured porcine kidney cell line PK-15. This analysis has revealed six novel transcripts and eight transcript isoforms, including five length and three splice variants. We were able to establish whether a deletion in a transcript is the result of the splicing of mRNAs or of genomic deletion in one of the PERV clones. Additionally, we re-annotated the formerly identified RNA molecules. Our analysis revealed a higher complexity of PERV transcriptome than it was earlier believed.

Introduction

The porcine endogenous retrovirus (PERV) is a C-type gammaretrovirus of swine (Todaro *et al.* 1974). The two polytropic subtypes: PERV-A and PERV-B infect cells of several species, including humans, while the ecotropic PERV-C infects only pig cells (Czauderna *et al.* 2000). The PK-15 cell line harbors 10 to 20 copies of at least two different clones of the PERV genome (Akiyoshi *et al.* 1998). The DNA homologous to PERV-PK (GeneBank accession: AJ293656) produces a full length

* Corresponding author

Name: Zsolt Boldogkői

Postal address: University of Szeged, Department of Medical Biology, Somogyi str. 4 Szeged, H-6720, Hungary

Tel.: +36/62-545109

Fax: +36/62-545131

E-mail: boldogkoi.zsolt@med.u-szeged.hu

8.3 kb long transcript and a subgenomic 3.1 kb long spliced transcript. This clone expresses all of the retroviral genes (*gag*, *pro/pol* and *env*), which enable the virus to productively infect the cells. The genome homologous to PK15-ERV (GeneBank accession: AF038601) has two deleted genomic regions affecting the RNA polymerase and envelope production of the virus, however this viral clone produces RNA molecules (Czauderna *et al.* 2000; Krach *et al.* 2001). The pig has been under consideration to possibly be utilized in the future as an organ donor organism (Wu *et al.* 2017). Therefore, PERV may represent a possible health hazard in xenotransplantations, as it is present in relatively high copy number in the genome of pigs. The elimination of PERV from the pig cell by specific pathogen-free breeding has been reported to be impossible, (Denner and Tönjes 2012), although some progress has been achieved recently by using CRISPR/Cas9-mediated excision of the provirus in cultured porcine kidney (PK-15) cells (Feng *et al.* 2015).

Whole transcriptome studies have become indispensable for understanding the complexity of genetic regulation. Short-read sequencing has become a commonly applied approach for the structural and functional annotation of transcriptomes (Mortazavi *et al.* 2008; Wang, Gerstein and Snyder 2009; Djebali *et al.* 2012). However, this technique is not optimal for the *de novo* characterization of the transcriptome, since it is unable to identify alternatively transcribed and processed transcripts and to distinguish between transcript isoforms, including splice and length variants. Both the Pacific Biosciences (PacBio) RS II platform and Oxford Nanopore Technologies (ONT) MinION platform are capable of reading long DNA stretches in a single sequencing run (Sharon *et al.* 2013; Laver *et al.* 2015). Although, short read technologies produce higher coverage than those of long-read sequencing, longer reads are much easier to process computationally and to interpret analytically since they map more unambiguously to a reference sequence and the larger the sequencing reads, the easier is to put the sequences together. Using longer reads, more overlaps can be identified between reads. In our former publications, we have demonstrated that PacBio long-read sequencing is able to reveal a hidden complexity of the transcriptional landscape of pseudorabies virus (PRV; (Tombácz *et al.* 2016), although the PRV transcripts have been formerly determined by using an Illumina short-read platform (Oláh *et al.* 2015). The PacBio long-read technique has also been successfully applied for the quantitation of dynamic PRV transcriptome (Tombácz *et al.* 2017). Both the PacBio isoform sequencing (Iso-Seq) library preparation and the ONT (1D strand switching cDNA by ligation) methods are able to determine the 5'-ends of the transcripts with base pair precision. PacBio sequencing has the important advantage over other methods in that it does not produce systematic errors and any that arise are therefore easily corrected thanks to its high consensus accuracy (Miyamoto *et al.* 2014). On the other hand the ONT sequencing technique does not have a skew towards a specific read length, thus it's more likely to sequence shorter and longer full length cDNAs than the PacBio platform (Weirather *et al.* 2017). However, the PacBio sequencing technique is more accurate than that of ONT, because if any random errors occur in the raw reads, they are easily corrected thanks to the exceptionally high consensus accuracy of this platform.

In this study, our aim was to reevaluate the PERV transcriptome with a multiplatform approach that included Illumina, PacBio and ONT cDNA sequencing, as well as regular PCR for validation purposes. We used oligo$(dT)_{20}$ primer-based sequencing for ONT, oligo$(dT)_{20}$ and random primer-based sequencing for both the PacBio and Illumina platforms. Additionally, we also applied both amplified and non-amplified PacBio Iso-Seq techniques.

Materials and methods

Cells and viruses

Immortalized porcine kidney cells of the cell line PK-15 (ATCC CCL-33) hosting PERV isolate Szeged were maintained in Dulbecco's modified Eagle's medium (Gibco Invitrogen) supplemented with 5% fetal bovine serum (Gibco, Invitrogen) and 80 μg mL-1 gentamycin (Gibco, Invitrogen) at 37°C, under 5% CO_2. Three freeze/thaw cycles were used, and the sample was then centrifuged at 10,000x g for 15min.

RNA purification

Total RNA isolation

Total RNA was isolated from the PK-15 cells with the Nucleospin RNA kit (Macherey-Nagel), according to the kit manual. Contaminating DNA was removed by on-column RNase free rDNase treatment (supplied by the kit), and the purified sample was further treated by TURBO DNA-free™ Kit (Life Technologies) to eliminate potential residual DNA contamination. RNA concentration was measured using the Qubit 2.0 and the RNA BR Assay Kit (Life Technologies). The RNA integrity was checked with Agilent 2100 Bioanalyzer. The RNA sample was stored at -80°C until use.

Isolation of polyadenylated RNAs

Isolation of the poly$(A)^+$ RNA fraction was carried out by the Spin-Column Protocol of the Oligotex mRNA Mini Kit (Qiagen). The RNA quantity was measured by Qubit RNA HS Assay Kit (Life Technologies).

Illumina HiScanSQ sequencing

Two library preparation methods were carried out for Illumina sequencing. Libraries were constructed for paired-end 100 bp sequencing using the Illumina ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre, Madison, WY USA) which includes a random-primed cDNA synthesis reaction step. For the poly(A)-sequencing, a single-end library was prepared by using anchored adaptor-primer oligonucleotides with a $(VN)T_{20}$ primer sequence. FastQC v0.10.1. was used for the quality assessment of raw read files.

PacBio RSII isoform sequencing

Three different library preparation approaches were used for SMRTbell library preparation.

The Non-amplified method Polyadenylated RNA samples were converted to cDNAs with SuperScript Double-Stranded cDNA Synthesis Kit (Life Technologies; the included SuperScript II was changed to SuperScript III enzyme). Anchored Oligo$(dT)_{20}$ primers (Life Technologies) were used for the reverse transcription reactions. Double-stranded cDNAs were quantified with the Qubit HS dsDNA Assay Kit (Life Technologies). SMRTbell libraries were prepared from cDNAs with PacBio DNA Template Prep Kit 1.0, following the *"Very Low (10 ng) Input 2 kb Template Preparation and Sequencing with Carrier DNA"* protocol. Agilent Technologies 2100 Bioanalyzer was used to determine the quality of the SMRTbell templates. DNA polymerase binding kit XL 1.0 and v2 sequencing primers were used for annealing and polymerase binding. The polymerase-template complexes were bound to MagBeads using the PacBio's MagBead Binding Kit. Sequencing was performed on the PacBio RSII platform with P5-C3 reagents. Movie lengths were 180 min (one movie was recorded for each SMRT cell).

Library preparation approaches from amplified cDNAs The following PacBio protocols were used for the cDNA production: Isoform Sequencing (Iso-Seq) using the SMARTer PCR cDNA Synthesis Kit (Clontech, as recommended by the manufacturer's) and No Size Selection or the Manual Agarose-gel Size Selection. Single-stranded cDNAs were synthesized from the polyA+ RNAs by using 3' SMART® CDS Primer II A (included in the Clontech kit) or adapter-linked GC-rich random primers.

The first-strand cDNAs were amplified by PCR, using the SMARTer PCR cDNA Synthesis Kit (Clontech) and the KAPA HiFi Enzyme (Kapa Biosystems) following the PacBio protocol. A 500ng

of cDNA sample was used for the SMRTbell library preparation, using the PacBio DNA Template Prep Kit 1.0.

DNA/Polymerase Binding Kit P6 kit was used for the production of the polymerase/template complexes. Sequencing was carried out on an RS II sequencer with DNA Sequencing Reagent 4.0 (P/N 100-356-200). The movie lengths were 240min (one movie was recorded for each SMRT cell).

Oxford Nanopore MinION sequencing

Libraries were prepared using the SQK-LSK108 Ligation Sequencing kit (Oxford Nanopore Technologies) applying the 1D strand switching cDNA by ligation protocol. Briefly: (ss)cDNA synthesis was carried out using SuperScript IV Reverse Transcriptase (Invitrogen/ Thermo Fisher Scientific) and an anchored adapter-primer with (VN)T_{20} nucleotides. A 5' adapter sequence with three O-methyl-guanine RNA bases was added for strand switching. PCR was carried out using the primers supplied in the kit and KapaHiFi high fidelity DNA polymerase (Kapa Biosystems). End repair was conducted using NEBNext End repair / dA-tailing Module (New England Biolabs) followed by adapter ligation using adapters supplied in the kit and NEB Blunt/TA Ligase Master Mix (New England Biolabs). CDNA was purified between each step using Agencourt AMPure XP magnetic beads (Beckman Coulter) and the sample concentration was determined using a Qubit 2.0 Fluorometer through the use of the Qubit (ds)DNA HS Assay Kit (Life Technologies/Thermo Fisher Scientific). Libraries were loaded on R9.5 SpotON Flow Cells, base calling was performed using Albacore v1.2.6.

PCR analysis

The putative novel and the previously described transcripts isoforms were validated by PCR analysis. CDNAs were created by reverse transcription with SuperScript IV Reverse Transcriptase (Life Technologies) following the manufacturer's recommendations. Samples were amplified by using the Applied Biosystem's Veriti Thermal Cycler with KAPA HiFi PCR Kit (KAPA Biosystems) according to the manufacturer's recommendations. The running conditions were as follows: 3 min at 95°C for initial denaturation, followed by 35 cycles at 98°C for 20 s (denaturation), 63°C for 20s (annealing), and at 72°C for 2min (extension). Final elongation was set at 72°C for 5 min. The primers used in this study are listed in (Table 1).

Data analysis and visualization

Reads from the Illumina sequencing were aligned with Bowtie 2 (Langmead and Salzberg 2012), while reads from PacBio and ONT sequencing with GMAP mapper (Wu and Watanabe 2005) to the host genome (Sus scrofa assembly: Sscrofa10.2) and to the genome of PERV isolate Szeged (GeneBank accession number: KY484771), which had been previously sequenced and aligned by our group. For visualization of mapped reads we used IGV (Thorvaldsdóttir, Robinson and Mesirov 2013). Poly(A) signals were predicted *in silico* by querying putative signal motifs to the 3' ends of the transcripts. TATA boxes were predicted using JASPAR POLII database (Mathelier *et al.* 2016) and FIMO (Find Individual Motif Occurrences) software (Grant, Bailey and Noble 2011) .

Results

Analysis of PERV transcriptome using multi-platform techniques

The Illumina HiScanSQ sequencing with random hexamer primers yielded 494,638 of 100-bp long reads with an average genome coverage of 5,703, while the polyadenylation sequencing (PA-Seq) method resulted in 99,499 of 50 nucleotide (nt)-long reads with an average coverage of 573, and a Reads Per Kilobase Million (RPKM) value of 0.099499. The PacBio long read sequencing yielded a total of 17,544 reads, and an average genome coverage of 3,238. The average length of the ROIs was 1,555 nts. The Oxford Nanopore Technologies MinION sequencing yielded 7,370 reads with an average read length of 1,512 nts and average genome coverage of 1,285.

Determination of the 5'- and 3'-ends of PERV transcripts

The 5' and the 3'-ends of the already described transcripts map to nucleotide 399 and nucleotide 8,625 respectively on the PERV-Szeged genome (Fig. 1, white arrow-rectangles). The upstream TATA-box was predicted *in silico* at genomic position 371-385, while the Poly(A) signal (PAS) of these transcripts was predicted to nt 8,596-8,601 (Table 2). Akiyoshi and colleagues described a 7.5 kilobase pair (kb)-long transcript using Northern blot analysis (Akiyoshi *et al.* 1998). We identified this as a 7,339 bp-long transcript mapping to 399 and 8,625 and have named it *gpe-d2*. Four of the six novel transcripts are located on the genome as follows: *inltr* mapped from nt 1 to 485, *gpe0-d* mapped from nt 1 to nt 8,625, *env-d* and *gpe-d* mapped from nt 399 to nt 8,625. The remaining two transcripts are truncated versions of *env*, and are located as follows: *env1.5* mapped from nt 7,248 to nt 8,625, *env1.3* mapped from nt 7,623 to nt 8,625 (Fig. 1. dark blue arrow-rectangles).

Identification of novel length-isoforms with PacBio sequencing

We have found two alternative transcript end sites (TES) one at positions 6,217 and the other at position 6,225, with the former being three times more abundant than the latter. PASs were predicted *in silico* to nt 6,198-6,203 (Table 2). Both of the previously described transcripts were found to be terminated alternatively in these positions. We designated these transcripts *gpe-AT* and *env-AT*. Two more transcripts from the deleted genomic copies named *gpe-d-AT* and *env-d-AT* were also found to be terminated at these positions. A 5' UTR length isoform of the truncated *env1.3* transcript designated *env1.3-L* is located between nts 7,563 and 8,625.

Determination of novel splice variants

We were able to identify three novel splice variants. *Env-d-1* has a 1,696 nt-long intron spanning from 6,509 to 8,205 (Fig. 2. lane E band 3). *Env-d-2* has an 1,849 nt-long intron having the same splice donor site as *env-d-1*, but a different acceptor site at position 8,358 (Fig. 2. lane E band 1). *Gpe-d2-1* has a 1,719-nt long intron with a donor site at 6,509 nt and an acceptor site at 8,329 nt (Fig. 2, lane E band 2).

Identification of novel PERV transcripts

We identified six novel transcripts and termed these *inltr, gpe0-d, env-d, gpe-d, env1.5* and *env1.3* (Fig. 1, dark blue arrow-rectangles). *Inltr* was detected in relatively low abundance using PacBio (0.6% of the reads), ONT (0.96% of the reads) and Illumina sequencing, The PAS of this transcript is predicted to nt 456-461. *In silico* analysis predicted a 90-bp long ORF encompassing 256 to 345 nt, which might encode a short, 30 amino acid long peptide. *Gpe0-d* is transcribed from a clone with a large deletion encompassing the region between nts 256-8,098, while *env-d* and *gpe-d* are transcribed from a genomic region with a deletion between nts 6,030-6,108. The two truncated versions of *env* incorporate two shorter in-frame ORFs. *Env1.5* harbors a 684 nt long, while *env1.3* a 276 nt ORF which may result in an N-terminally truncated env protein.

Determination of the deletions of PERV

Mapping the sequencing reads to the genome of PERV-Szeged uncovered several intronic regions. These could be the results of either splicing or deleted genomic segments. PCR analysis revealed multiple deleted regions on the viral genome (Fig. 1., dashed lines). Based on this data, we suggest the existence of at least four different PERV clones in the genome of PK-15 cells. The first clone harbors no deletion; the second one is deleted between 256-8,098 nt (Fig. 2, lane A and A'); the third is deleted between nt 6,030 – 6,108 (Fig. 2, lane D and D'), while the forth clone has a deletion between nt 3,932 – 4,740 (Fig. 2, lane C and C') and 6,030 – 6,108 (Fig. 2, lane D and D').

Discussion

Short-read sequencing techniques – despite their widespread use for structural annotation of transcriptomes – are suboptimal for example for the identification of alternatively transcribed and

spliced transcripts (Mortazavi *et al.* 2008) and transcript isoforms. Long-read sequencing however has proven to be an excellent platform for the identification of splice and length transcript variants in human (Sharon *et al.* 2013) and herpesviruses (Tombácz *et al.* 2016, 2017).

In this study, we applied a massively parallel long-read sequencing platform to characterize the transcriptome of PERV in PK-15 cells. We identified six novel transcripts, five length isoforms and three splice variants. The relative low amounts of some of these transcripts and their overlapping nature makes their identification difficult in gel-based assays and PCR. Additionally, we were also able to pinpoint four deleterious clones of PERV-Szeged using PCR.

Our investigation has succeeded in uncovering a complex transcriptional landscape in PERV, and to characterize the deletions of the provirus clones in the host genome.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets generated and/or analysed during the current study are available in the Sequence Read Archive database accessible under accession: PRJNA381012.

Competing interests

The authors declare no conflict of interests. The founding sponsors had no role in the design of the study, in the collection, analyses, or interpretation of data, nor in the writing of the manuscript or in the decision to publish the results.

Funding

This work was supported by the following: National Institutes of Health (NIH) Centers of Excellence in Genomic Science (CEGS) - Center for Personal Dynamic Regulomes: [grant number 5P50HG00773502 to MS]; TAMOP-Social Renewal Operational Programme: [grant number TAMOP-4.2.6-14/1 -288 to ZBo]; Bolyai Janos Scholarship of the Hungarian Academy of Sciences: [grant number 2015-18 to DT], and Swiss-Hungarian Cooperation Programme [grant number SH/7/2/8 to ZBo].

Authors' contributions

NM carried out the PCR experiments, analysed the data, participated in the sequence alignment and drafted the manuscript. ZBa analysed the data and participated in the sequence alignment. DT

carried out the PacBio sequencing, participated in the design of the study and took part in drafting the manuscript. ZC propagated the cells, prepared the RNA, DNA and cDNA samples. AS participated in the sequence alignment. MS participated in the coordination of the study. ZBo conceived, designed and coordinated the study and wrote the manuscript. All authors have read and approved the final version of the manuscript.

Acknowledgements
We would like to thank Marianna Ábrahám and Csilla Magyarné Papdi (University of Szeged) for technical assistance.

References

Akiyoshi DE, Denaro M, Zhu H *et al.* Identification of a full-length cDNA for an endogenous retrovirus of miniature swine. *J Virol* 1998;**72**:4503–7.

Czauderna F, Fischer N, Boller K *et al.* Establishment and characterization of molecular clones of porcine endogenous retroviruses replicating on human cells. *J Virol* 2000;**74**:4028–38.

Denner J, Tönjes RR. Infection barriers to successful xenotransplantation focusing on porcine endogenous retroviruses. *Clin Microbiol Rev* 2012;**25**:318–43.

Djebali S, Davis CA, Merkel A *et al.* Landscape of transcription in human cells. *Nature* 2012;**489**:101–8.

Feng W, Dai Y, Mou L *et al.* The potential of the combination of CRISPR/Cas9 and pluripotent stem cells to provide human organs from chimaeric pigs. *Int J Mol Sci* 2015;**16**:6545–56.

Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;**27**:1017–8.

Krach U, Fischer N, Czauderna F *et al.* Comparison of replication-competent molecular clones of porcine endogenous retrovirus class A and class B derived from pig and human cells. *J Virol* 2001;**75**:5465–72.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.

Laver T, Harrison J, O'Neill PA *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* 2015;**3**:1–8.

Mathelier A, Fornes O, Arenillas DJ *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2016;**44**:D110–5.

Miyamoto M, Motooka D, Gotoh K *et al.* Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics* 2014;**15**:699.

Mortazavi A, Williams BA, McCue K *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.

Oláh P, Tombácz D, Póka N *et al.* Characterization of pseudorabies virus transcriptome by Illumina sequencing. *BMC Microbiol* 2015;**15**:130.

Sharon D, Tilgner H, Grubert F *et al.* A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* 2013;**31**:1009–14.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;**14**:178–92.

Todaro GJ, Benveniste RE, Lieber MM *et al.* Characterization of a type C virus released from the porcine cell line PK(15). *Virology* 1974;**58**:65–74.

Tombácz D, Balázs Z, Csabai Z *et al.* Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. *Sci Rep* 2017;**7**:43751.

Tombácz D, Csabai Z, Oláh P *et al.* Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. Banfield BW (ed.). *PLoS One* 2016;**11**:e0162868.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.

Weirather JL, de Cesare M, Wang Y *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 2017;**6**:100.

Wu J, Platero-Luengo A, Sakurai M *et al.* Interspecies Chimerism with Mammalian Pluripotent Stem Cells. *Cell* 2017;**168**:473–486.e15.

Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005;**21**:1859–75.

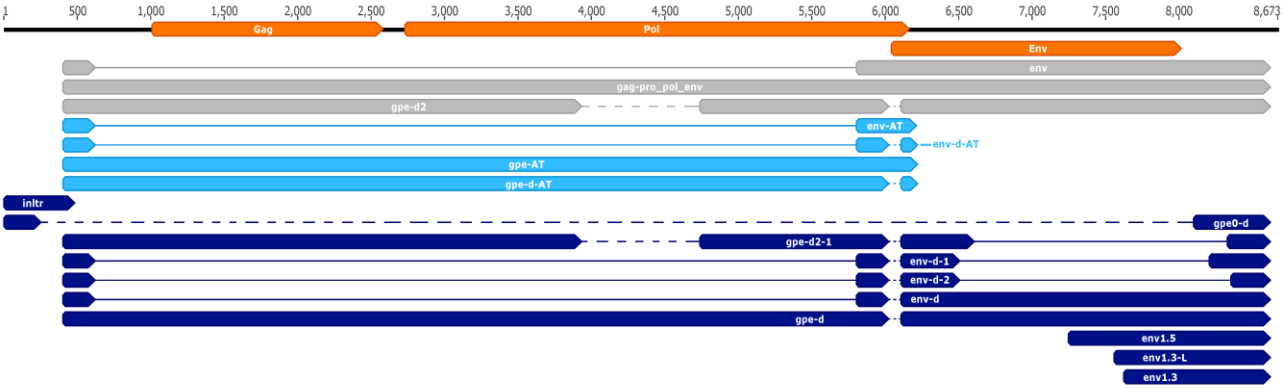

Fig. 1. Location of the already characterized and the novel transcripts on the PERV-Szeged genome. Arrow-rectangles in orange: ORFs, arrow-rectangles in gray: already known transcripts, arrow-rectangles in light blue: length variants, arrow-rectangles in dark blue: novel transcripts, novel splice isoforms. Dashed lines represent deletions of the genome while continuous lines represent introns.

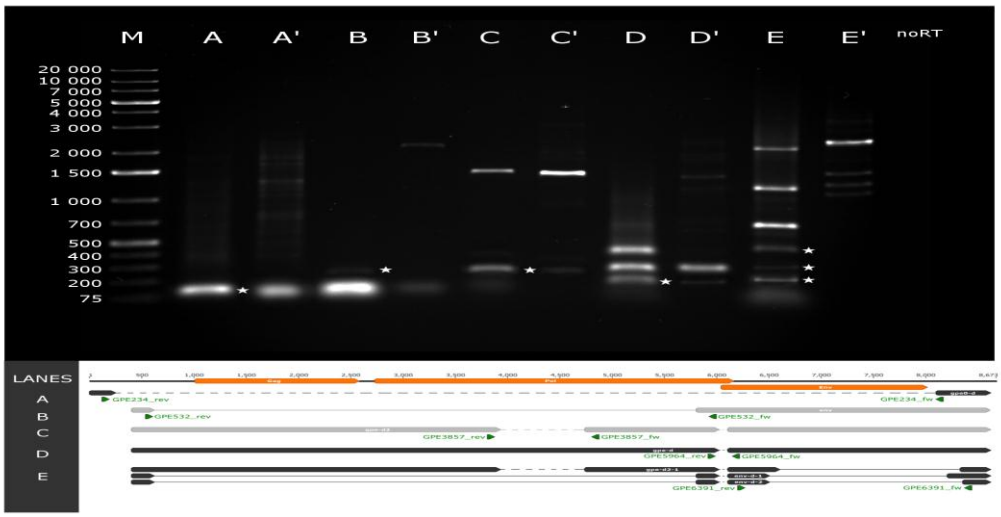

Fig. 2. Novel splice variants and transcripts of the PERV-Szeged clones. 1% agarose gel electrophoresis of the novel splice variants and new transcripts of the PERV-Szeged clones. Lanes missing an apostrophe were loaded with cDNA products from RT-PCR while lanes with the same letter and an apostrophe were loaded with PCR products of genomic DNA. Lane M was loaded with

GeneRuler 1kb Plus DNA Ladder (Thermo Fisher Scientific). Staining was performed with GelRed (Biotium). Lane no-RT was loaded with no-RT control. On lanes A, A', C, C' and D, D' bands marked with a star represent new transcripts of the PERV-Szeged clones with an amplicon length of 90, 219 and 216 bp respectively. On lane B, B' bands marked with a star represent the already known *env* transcript. On lane E, E' bands marked represent three new splice variants with an amplicon length of 274, 304 and 427 bp. The size of each band of the 1kb Plus DNA Ladder is presented to the right of it's lane. Green arrows represent the location of the PCR primers.

Table 1. Novel transcripts. **TSS: transcript start site, PA: poly(A), TES: transcript end site; positions separated with a slash represent separate entries; asterisks (*) mark a deletion.**

Tr. name	TATA-box	TSS	Exon end	Exon start	PA signal	TES
gpe0-d	-	1	255*	8,098*	8,596-8,601	8625
inltr	-	1	-	-	456-461	485
env-AT	371-385	399	623	5,807	6,198-6,203	6,217/6,225
gpe-d	371-385	399	6,029*	6,109*	6,198-6,203	6,217/6,225
env-d-AT	371-385	399	623/6,029*	5,807/6,109*	6,198-6,203	6217/6,225
gpe-d2-1	371-385	399	3,931*/6,029*/6,509	4,741*/6,109*/8,328	8,596-8,601	8,625
env-d-1	371-385	399	623/6,029*/6,509	5,807/6,109*/8,205	8,596-8,601	8,625
env-d-2	371-385	399	623/6,029*/6,509	5,807/6,109*/8,358	8,596-8,601	8,625
gpe-AT	371-385	399	-	-	6,198-6,203	6,217/6,225
env	371-385	399	623	5,807	8,596-8,601	8,625
gpe-d-AT	371-385	399	6,029*	6,109*	8,596-8,601	8,625
env-d	371-385	399	623/6,029*	5,807/6,109*	8,596-8,601	8,625
gpe-d2	371-385	399	3,931*/6,029	4,741*/6,109	8,596-8,601	8,625
env1.5	-	7248	-	-	8,596-8,601	8,625
env1.5	-	7563	-	-	8,596-8,601	8,625
env1.5	-	7623	-	-	8,596-8,601	8,625

Table 2. Primers used for PCR analysis. **The start positions mark the first nucleotide of the primer on the genome of PERV-Szeged**

Name	**Sequence**	**Start Position**
GPE234_fw	TGGCAGCCAGCAGGGTCTGG	8,167
GPE234_rev	GGACCTCCGGAGCTATTTTA	234
GPE532_fw	AACATAGACTGAATCTCCAA	5,935
GPE532_rev	ACGAGGGGGATTGTTCTTTT	532
GPE3857_fw	TCCCTCTAGATATGAGATCT	4,883
GPE3857_rev	GCTGGATTTTGCAGACTGTG	3,857
GPE5964_fw	GTTTATGGGAGTTCGGGCTG	6,204
GPE5964_rev	CTCGGTGGAAGGGACCTTAT	5,964
GPE6391_fw	ATGGGGTTCACAACAAAGCC	8,514
GPE6391_rev	CAGGACCCCCAAATAATGAA	6,391